UNIVERSITY OF HONG KONG
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE

Topics for STAT2318 Directed Studies in Statistics (6 credits)
(2013 – 2014)

1. Approximations for ruin probabilities in insurance risk theory

The (infinite-time) ruin probability can be regarded as a measure of the risk associated to an insurance company. Unfortunately, even in the simplest case of the compound Poisson risk model, closed-form expressions only exist under certain claim severity distributions. This project aims at studying various approximation methods for ruin probability, which include, for example, De Vylder approximation, Tijms' approximation and Dickson-Waters' discretization method. Finite-time ruin probabilities will also be discussed. Strong knowledge in stochastic processes and computational skills will be required. The student is assumed to be familiar with software packages such as Mathematica, Maple or Matlab.

Supervisor: **Dr. Eric C.K. Cheung**, Department of Statistics and Actuarial Science (MW502J)
(eckc@hku.hk)

2. Copulas in risk management

Copulas are functions that join multivariate distribution functions to their one-dimensional marginal distribution functions. The student who takes this project is expected to study the basic theory of copula and some of its applications in risk management. All the related literature will be provided.

Supervisor: **Dr. K.C. Cheung**, Department of Statistics and Actuarial Science (MW522)
(kccg@hku.hk)

3. H shares and A shares

Many Chinese companies in China are dual-listed in Hong Kong and China (Shanghai or Shenzhen) by issuing H shares and A shares respectively, with price discrepancies having been found between them. The student who takes this project is expected to study the relationship between the movements of H shares and A shares, taking into account various economic factors.

Requirement: Knowledge of financial markets and SAS programming.

Supervisor: **Dr. K.S. Chong**, Department of Statistics and Actuarial Science (MW504)
(kschong@hku.hk)

4. Analysis of Card Counting Strategies in Casino Games

Card counting is the process of tracking the live cards played in a card game in order to determine when the deck is in favor of the player and hence when the player should increase the bet size. Since its development in the 1950's, the card counting strategy has been known to be capable of effectively increasing the player's edge in some casino games, especially blackjack. In this project, students will be asked to perform mathematical analysis on some selected casino games and explore different existing card counting strategies using Monte Carlo simulations. Students are also encouraged to design new card counting systems; or suggest modifications to the game rules to prevent card counters from profiting.

Requirement: Knowledge in programming language like FORTRAN or C+ is a MUST.

Supervisor: **Dr. Y.K. Chung**, Department of Statistics and Actuarial Science (MW504)
(yukchung@hku.hk)

5. Analysis of population census data

The Hong Kong Population Census was conducted in 2011. The census data provide a lot of information on the social and economic situation in Hong Kong. In this project, the student will analyse the 5% sample data set for the 2011 Population Census.

Requirement: Initiative and good knowledge about the social and economical situation in Hong Kong.

Supervisor: **Prof. W.K. Fung**, Department of Statistics and Actuarial Science (MW523)
(wingfung@hku.hk)

6. EM Algorithms for ML Factor Analysis

The general theory of EM algorithms proves that each iteration of EM increases the likelihood and also that if an instance of the algorithm converges, it converges to a (local) maximum of the likelihood. An advantage of EM algorithms, such as those for factor analysis, is that each iteration is simple to program and computationally inexpensive. This project is to explore the application of EM algorithms for maximum likelihood factor analysis.

Requirement: Knowledge of programming in SAS/IML or other programming languages is essential.

Supervisor: **Dr. C.W. Kwan**, Department of Statistics and Actuarial Science (MW508)
(cwkwan@hku.hk)

7. Analysis of correlated zero-inflated count data

In many medical and public health investigations, the count data encountered often exhibit an excess of zeros, and very frequently this type of data are collected on clusters of subjects or by repeated measurements on each subject. For example, in the analysis of medical expenditure, members in the same family may exhibit some correlation possibly due to housing locality, genetic predisposition, similar dietary and living habit. Ignoring such correlation may lead to misleading statistical inference. This project will survey the models and methods in the literature and apply them to a real data set.

Requirement: Knowledge in programming language like FORTRAN or C++.

Supervisor: **Dr. Eddy K.F. Lam**, Department of Statistics and Actuarial Science (MW519)
(hrntlkf@hku.hk)

8. Statistical analysis of "Fame"

About a decade ago, researchers at UCLA conducted an interesting study on the statistical relationship between "fame" and "achievement" of World War I pilots. They employed a novel definition of "fame", measured as the number of hits a person's name garners in Google search. The results are finally published in Simkin and Roychowdhury (2006) in Journal of Mathematical Sociology. This project explores the statistical issues involved in studies of this kind and conducts statistical analyses of real-life data of the student's own choice.

Supervisor: **Prof. Stephen M.S. Lee**, Department of Statistics and Actuarial Science (MW528)
(smslee@hku.hk)

9. Bootstrap approximation in time series modeling

The traditional time series modeling and further inference are based on the normality assumption or large enough sample size. In the real applications, the normality may be broken and the results may not be accurate for the moderate or small sample sizes. The bootstrap is a computer-intensive method, and the information in the real data is repeatedly used. Hence it may provide more accurate results. This project hopefully can train students for some bootstrap methods to dependent data, and some knowledge of computer languages such as FORTRAN or C is required since a little more computation will be involved.

Supervisor: **Dr. G. Li**, Department of Statistics and Actuarial Science (MW502H)
(gdli@hku.hk)

10. Technical Pattern Identification of Stock Market

As a technical analyst of stock market, identification of chart pattern is an essential procedure before any forecasting procedure of stock price. In this study, chart patterns of selected stocks and markets will be investigated. Since pattern identification may not be a straightforward procedure throughout the process, intensive application of statistical or data mining techniques may be considered.

Requirement: Knowledge of financial market and R programming

Supervisor: **Dr. Gilbert C.S. Lui,** Department of Statistics and Actuarial Science (MW506)
(csglui@hku.hk)

11. Modeling of Risk Factors for Solvency Calculations of Life Insurance Companies

In order to fulfill the requirements of Solvency II, European life insurance companies have to develop internal models for calculations of the asset and liability values. However, these models are usually based on a complete valuation of the entire portfolio of the company and can be very time consuming when the company wants to test the impact of various risk factors and control variables on the solvency position. In this project, the student will analyze the risk factors affecting the solvency position of a typical insurance company and develop appropriate simple models for the solvency calculations. The student is expected to collect information relating to solvency calculations and experience data of various risk factors. The end result should be a workable model that can produce solvency calculations for a typical insurance company to perform analysis of the different levels of risk factors and resulting solvency requirements under various scenarios.

Supervisor: **Dr. Louis F.K. Ng**, Department of Statistics and Actuarial Science (MW505)
(flouisng@hku.hk)

12. Bayesian inference using MCMC sampling

In this project, the student shall learn the Bayesian inference using MCMC sampling methods, with some computer software, WINBUGS, SAS or R code.

Supervisor: **Prof. K.W. Ng**, Department of Statistics and Actuarial Science (MW525)
(kaing@hku.hk)

13. Maximum likelihood estimates of parameters in multivariate zero-inflated Poisson Models

Zero-truncated and zero-inflated count data often occur in areas such that public health, epidemiology, medicine, sociology, psychology, engineering, agriculture, and ecology. Zero-truncated count data mean that the response variable cannot have a value of zero. A typical example from medical literature is the duration patients are in hospital. Zero-inflated count data mean that the response variable contains more zeros than expected. Yip (1988) used the inflated Poisson distribution to model the number of insects per leaf. Heilbron (1989) proposed zero-altered Poisson and negative binomial regression models and applied them to study high-risk human behavior. Lambert (1992) used the ZIP regression model to study covariate effects with an example from modeling experimental results of soldering defects on printed wiring boards. Gupta, Gupta and Tripathi (1995, 1996) proposed zero-adjusted discrete models including the zero-inflated modified power series distributions.

In this project, the admitted candidates are expected to (i) review literature on univariate zero-truncated/inflated/altered Poisson, generalized Poisson, negative binomial, binomial, and beta-binomial models with and without covariates, and/or (ii) design a feasible optimization algorithm for finding maximum likelihood estimates of parameters in multivariate zero-inflated Poisson models, and (iii) to analyze two real data sets via R or SAS program.

Requirement: Knowledge of LATEX technique, optimization, and R programming or SAS/IML

Supervisor: **Dr. Gary G. Tian**, Department of Statistics and Actuarial Science (MW520)
(gltian@hku.hk)

14. Comparison of Several Free Statistical Packages

This project aims to compare various free or even open-source statistical software packages, thereby providing students with opportunities to perform statistical computations with different resources. In this project, students should gain extensive experience in programming with several free statistical packages such as R, Sage, Scilab, SciPy, etc. Some common statistical analyses and simulation studies are to be carried out by using those packages and students are required to learn the way to code the programs for several statistical problems. Students taking this project are expected to be self-motivated in learning programming skills.

Supervisor: **Dr. K.P. Wat**, Department of Statistics and Actuarial Science (MW529)
(watkp@hku.hk)

15. Investigation of Non-normality in a Simple Errors-in-variables Model

In a classical linear regression model, it is usually assumed that the predictive variable is not subject to any kind of random error. However, it is not always true in many applications. In addition, it is also a common practice to assume that the error in the regression model is normally distributed. Unfortunately, we may often find that most real data sets do not really exhibit such nice properties. In this project, student will investigate the non-normality situation where the errors in a regression model exist. Computer programming skill is required.

Requirement: Strong knowledge in computer programming and statistical simulation technique is a must.

Supervisor: **Dr. Raymond W.L. Wong**, Department of Statistics & Actuarial Science (MW511) (rwong@hku.hk)

16. Risk theoretic applications with a class of mixed Erlangs

It is known that a variety of distributions are of mixed Erlang type, in which case computational formula exists for many quantities of interests in risk theory. In this project, the student studies distributional properties of the class of Erlang mixtures as well as various risk theoretic applications including analysis of insurer's surplus process and discounted aggregate claims. In particular, the student is assumed to possess strong computational skills such as Mathematica, Matlab or Maple, also statistical background related to the EM algorithm.

Supervisor: **Dr. J.K. Woo**, Department of Statistics & Actuarial Science (MW530) (jkwoo@hku.hk)

17. Modelling Mean and Dispersion of Data of the Exponential Family

One way to model Non-Gaussian data without transformation is the application of Generalized Linear Model (GLM) proposed by Nelder and Wedderburn (1972) as long as the underlying distribution of the response variable belongs to the exponential family. In the theory of GLM, dispersion of the response variable is treated as constant and can be estimated by the sample. In this project, students are expected to investigate the appropriateness of assuming constant dispersion in practical analysis, especially if the response variable is Poisson or Binomial distributed, and the alternative of modeling mean and dispersion jointly as suggested by Smyth and Verbyla (1999).

Supervisor: **Dr. Karl K.Y. Wu**, Department of Statistics & Actuarial Science (MW508) (karlwu@hku.hk)

18. Insurance Risk

In this project, the student will study the following topics: Probability distributions, utility theory, principles of premium calculation, the individual and collective risk models, and basic ruin theory.

Supervisor: **Prof. H. Yang**, Department of Statistics & Actuarial Science (MW526)
(hlyang@hku.hk)

19. Analysis of large data sets: new tools from random matrix theory

Large data sets refer to data where the number of variables, or data dimension say *p*, is large compared to the sample size, say *n*. Modern statistical problems involve frequently such large data sets from various fields like genomic data analysis, financial portfolio optimization or design of wireless communication networks. For example in a genomic micro-array, *p* is several thousands and *n* several hundreds. Classical multivariate statistical tools dramatically fail to analyse these large data sets: either there are not applicable any more or lack efficiency.

New tools have emerged recently from the theory of random matrices. Most of them are based on the distribution of eigenvalues of sample covariance matrices which are computable from the data. Classical tools like Hotelling T^2 (tests on the mean), testing of generalized linear hypothesis (for regression or MANOVA) have been corrected or adapted to cope with large data sets. The theory behind is appealing and applications to large-dimensional data analysis are significant.

In this project, the student will i) learn some fundamental theorems from the theory of random matrices; ii) learn some new statistical tools developed recently; iii) start some own thinking about unsolved problems or perform some simulation experiments in order to get a deeper understanding of these results. I have included below an expository paper on the subject for a first introduction.

First reading:
Z.D. BAI (2005). High dimensional data analysis. *COSMOS*, Vol. 1, No. 1, 17–27.
(downloadable from: http://web.hku.hk/~jeffyao/Bai-cosmos-05.pdf)

Supervisor: **Dr. Jeff J. Yao**, Department of Statistics & Actuarial Science (MW502G)
(jeffyao@hku.hk)

20. Bayesian Hierarchical Modeling and Dose Finding with Longitudinal Data

Longitudinal data are common in clinical trials for Alzheimer's disease. In a phase II trial, the goal is to identify one or several doses that may have disease modification effects, which would be moved forward to phase III clinical trials for confirmative testing. Students will develop Bayesian hierarchical modeling for longitudinal measurements, and incorporate the slopes of disease deterioration into decision making. In the Bayesian adaptive framework, any arm that shows futility would be terminated earlier, so that the rest of patients would be allocated to the remaining arms. Students will conduct extensive simulation studies to demonstrate the performance of the Bayesian hierarchical modeling and dose finding method.

Requirement: R programming

Supervisor: **Dr. G. Yin**, Department of Statistics and Actuarial Science (MW502E)
(gyin@hku.hk)

21. The Effects of After-Hours Futures Trading

Since April 8, 2013, Hong Kong Exchanges and Clearing Limited (HKEx) introduced after-hours futures trading (AHFT) on two of its futures products: Hang Seng Index futures and H-shares Index futures. The AHFT aims to enable market participants to hedge or adjust their positions in response to market news and events during the European and US business days. Note that if there are fewer buyers and sellers to generate fluid markets, futures prices may tend to fluctuate more in the after-hours markets. We would like to study whether the after-hours futures trading would ultimately hurt or benefit investors. More specifically, we would study whether volatility in the after-hours markets increases or decreases.

Supervisor: **Dr. Philip L.H. Yu**, Department of Statistics and Actuarial Science (MW521)
(plhyu@hku.hk)

22. What has driven up CPI in Hong Kong?

Students registered in this project are going to collect various of variables that might be related to dynamic of CPI in Hong Kong, and build proper regression model to identify any possible/potential reasons/resources for the driving up of CPI in Hong Kong these years.

Supervisor: **Dr. Z. Zhang**, Department of Statistics and Actuarial Science (MW511)
(zhangz08@hku.hk)

******** END ********